Concentration of Measure Inequalities for Toeplitz Matrices with Applications

Borhan M. Sanandaji, Tyrone L. Vincent, and Michael B. Wakin

Abstract

We derive Concentration of Measure (CoM) inequalities for randomized Toeplitz matrices. These inequalities show that the norm of a high-dimensional signal mapped by a Toeplitz matrix to a low-dimensional space concentrates around its mean with a tail probability bound that decays exponentially in the dimension of the range space divided by a quantity which is a function of the signal. For the class of *sparse* signals, the introduced quantity is bounded by the sparsity level of the signal. However, we observe that this bound is highly pessimistic for most sparse signals and we show that if a random distribution is imposed on the non-zero entries of the signal, the typical value of the quantity is bounded by a term that scales logarithmically in the ambient dimension. As an application of the CoM inequalities, we consider Compressive Binary Detection (CBD).

Index Terms

Concentration of Measure Inequalities, Compressive Toeplitz Matrices, Compressive Sensing.

I. INTRODUCTION

A. Overview

Motivated to reduce the burdens of acquiring, transmitting, storing, and analyzing vast quantities of data, signal processing researchers have over the last few decades developed a variety of techniques for data compression and dimensionality reduction. Unfortunately, many of these techniques require a raw, high-dimensional data set to be acquired before its essential low-dimensional structure can be identified, extracted, and exploited. In contrast, what would be truly desirable are sensors/operators that require fewer raw measurements yet still capture the essential information in a data set. These operators can be called *compressive* in the sense that they act as mappings from a high-dimensional to a low-dimensional

space, e.g., $X : \mathbb{R}^N \to \mathbb{R}^M$ where $M < N$. Linear compressive operators correspond to matrices having fewer rows than columns. Although such matrices can have arbitrary/deterministic entries, *randomized* matrices (those with entries drawn from a random distribution) have attracted the attention of researchers due to their universality and ease of analysis. Utilizing such compressive operators to achieve *information-preserving embeddings* of high-dimensional (but compressible) data sets into low-dimensional spaces can drastically simplify the acquisition process, reduce the needed amount of storage space, and decrease the computational demands of data processing.

Concentration of Measure (CoM) inequalities are one of the leading techniques used in the theoretical analysis of randomized compressive linear operators [1]–[3]. These inequalities quantify how well a random matrix will preserve the norm of a high-dimensional signal when mapping it to a low-dimensional space. A typical CoM inequality takes the following form. For any fixed signal $\boldsymbol{a} \in \mathbb{R}^N$, and a suitable random $M \times N$ matrix X, the random variable $\|X\boldsymbol{a}\|_2^2$ will be highly concentrated around its expected value, $\mathbf{E}\left[\|X\boldsymbol{a}\|_2^2\right]$, with high probability. Formally, there exist constants c_1 and c_2 such that for any fixed $\boldsymbol{a} \in \mathbb{R}^N$,

$$\mathbf{P}\left\{\left|\|X\boldsymbol{a}\|_2^2 - \mathbf{E}\left[\|X\boldsymbol{a}\|_2^2\right]\right| \geq \epsilon \mathbf{E}\left[\|X\boldsymbol{a}\|_2^2\right]\right\} \leq c_1 e^{-c_2 M c_0(\epsilon)}, \tag{1}$$

where $c_0(\epsilon)$ is a positive constant that depends on $\epsilon \in (0, 1)$.

CoM inequalities for random operators have been shown to have important implications in signal processing and machine learning. One of the most prominent results in this area is the Johnson-Lindenstrauss (JL) lemma, which concerns embedding a finite set of points in a lower dimensional space using a distance preserving mapping [4]. Dasgupta et al. [2] and Achlioptas [3] showed how a CoM inequality of the form (1) could establish that with high probability, an independent and identically distributed (i.i.d.) random compressive operator $X \in \mathbb{R}^{M \times N}$ ($M < N$) provides a JL-embedding. Specifically, for a given $\epsilon \in (0, 1)$, for any fixed point set $Q \subseteq \mathbb{R}^N$,

$$(1 - \epsilon)\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \leq \|X\boldsymbol{a} - X\boldsymbol{b}\|_2^2 \leq (1 + \epsilon)\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \tag{2}$$

holds with high probability for all $\boldsymbol{a}$, $\boldsymbol{b} \in Q$ if $M = \mathcal{O}\left(\epsilon^{-2}\log\left(|Q|\right)\right)$. One of the other significant consequences of CoM inequalities is in the context of Compressive Sensing (CS) [5] and the Restricted Isometry Property (RIP) [6]. If a matrix X satisfies (2) for all pairs $\boldsymbol{a}$, $\boldsymbol{b}$ of K-sparse signals in $\mathbb{R}^N$, then X is said to satisfy the RIP of order $2K$ with isometry constant ϵ. Establishing the RIP of order $2K$ for a given compressive matrix X leads to understanding the number of measurements required to have exact recovery for any K-sparse signal $\boldsymbol{a} \in \mathbb{R}^N$. Baraniuk et al. [7] and Mendelson et al. [8] showed that CoM inequalities can be used to prove the RIP for random compressive matrices.

CoM inequalities have been well-studied and derived for *unstructured* random compressive matrices, populated with i.i.d. random entries [2], [3]. However, in many practical applications, measurement matrices possess a certain structure. In particular, when linear dynamical systems are involved, Toeplitz and circulant matrices appear due to the convolution process [9]–[12]. Specifically, consider the linear time-invariant (LTI) dynamical system with system finite impulse response $a = \{a_k\}_{k=1}^{N}$. Let $x = \{x_k\}_{k=1}^{N+M-1}$ be the applied input sequence. Then the corresponding output is calculated from the time-domain convolution of a and x. Supposing the x_k and a_k sequences are zero-padded from both sides, each output sample y_k can be written as

$$y_k = \sum_{j=1}^{N} a_j x_{k-j}. \tag{3}$$

If we keep only M consecutive observations of the system, $y = \{y_k\}_{k=N+1}^{N+M}$, then (3) can be written in matrix-vector multiplication format as

$$y = Xa, \tag{4}$$

where

$$X = \begin{bmatrix} x_N & x_{N-1} & \cdots & x_1 \\ x_{N+1} & x_N & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{N+M-1} & x_{N+M-2} & \cdots & x_M \end{bmatrix} \tag{5}$$

is an $M \times N$ Toeplitz matrix. If the entries of X are generated randomly, we say X is a randomized Toeplitz matrix. Other types of *structured* random matrices also arise when dynamical systems are involved. For example block-diagonal matrices appear in applications such as distributed sensing systems [13] and initial state estimation (observability) of linear systems [14].

In this paper, we consider compressive randomized Toeplitz matrices, derive CoM inequalities, and discuss their implications in applications such as sparse impulse response recovery [10], [15], [16]. We also consider the problem of detecting a deviation in a system's behavior. We show that by characterizing the deviation using a particular measure that appears in our CoM inequality, the detector performance can be correctly predicted.

B. Related Work

Compressive Toeplitz (and circulant) matrices have been previously studied in the context of CS [9]–[11], [15], [17], [18], with applications involving channel estimation, synthetic aperture radar, etc. Tropp

et al. [17] originally considered compressive Toeplitz matrices in an early CS paper that proposed an efficient measurement mechanism involving a Finite Impulse Response (FIR) filter with random taps. Motivated by applications related to sparse channel estimation, Bajwa et al. [18] studied such matrices more formally in the case where the matrix entries are drawn from a symmetric Bernoulli distribution. Later they extended this study to random matrices whose entries are bounded or Gaussian-distributed and showed that with high probability, $M \geq \mathcal{O}\left(K^2 \log\left(\frac{N}{K}\right)\right)$ measurements are sufficient to establish the RIP of order $2K$ for vectors sparse in the time domain [10], [15]. (It should be noted that the quadratic RIP result can also be achieved using other methods such as a coherence argument [9], [19].) Recently, using more complicated mathematical tools such as Dudley's inequality for chaos and generic chaining, Rauhut et al. [20] showed that with $M \geq \mathcal{O}\left(K^{1.5} \log\left(N\right)^{1.5}\right)$ measurements the RIP of order $2K$ will hold.[1] Note that these bounds compare to $M \geq \mathcal{O}\left(K \log\left(\frac{N}{K}\right)\right)$ measurements which are known to suffice when X is unstructured [7].

In this paper, we derive CoM inequalities for Toeplitz matrices and show how these inequalities reveal non-uniformity and signal-dependency of the mappings. As one consequence of these CoM inequalities, one could use them (along with standard covering number estimates) to prove the RIP for compressive Toeplitz matrices. Although the estimate of the required number of measurements would be quadratic in terms of sparsity (i.e., $M \sim K^2$) and fall short of the best known estimates described above, studying concentration inequalities for Toeplitz matrices is of its own interest and gives insight to other applications such as the binary detection problem.

There also exist CoM analyses for other types of structured matrices. For example, Park et al. [13] derived concentration bounds for two types of block diagonal compressive matrices, one in which the blocks along the diagonal are random and independent, and one in which the blocks are random but equal.[2] We subsequently extended these CoM results for block diagonal matrices to the observability matrices that arise in the analysis of linear dynamical systems [14].

---

[1]After the initial submission of our manuscript, in their very recent work, Krahmer et al. [21] showed that the minimal required number of measurements scales linearly with K, or formally $M \geq \mathcal{O}\left(K \log\left(K\right)^2 \log\left(N\right)^2\right)$ measurements are sufficient to establish the RIP of order $2K$. The recent linear RIP result confirms what is suggested by simulations.

[2]Shortly after our own development of CoM inequalities for compressive Toeplitz matrices (a preliminary version of Theorem 1 appeared in [12]), Yap and Rozell [22] showed that similar inequalities can be derived by extending the CoM results for block diagonal matrices. Our Theorem 2 and the associated discussion, however, is unique to this paper.

C. Contributions

In summary, we derive CoM inequalities for randomized Toeplitz matrices. The derived bounds in the inequalities are non-uniform and depend on a quantity which is a function of the signal. For the class of *sparse* signals, the introduced quantity is bounded by the sparsity level of the signal while if a random distribution is imposed on the non-zero entries of the signal, the typical value of the quantity is bounded by a term that scales logarithmically in the ambient dimension. As an application of the CoM inequalities, we consider Compressive Binary Detection (CBD).

II. MAIN RESULTS

In this paper, we derive CoM bounds for compressive Toeplitz matrices as given in (5) with entries $\{x_k\}_{k=1}^{N+M-1}$ drawn from an i.i.d. Gaussian random sequence. Our first main result, detailed in Theorem 1, states that the upper and lower tail probability bounds depend on the number of measurements M and on the eigenvalues of the covariance matrix of the vector $\boldsymbol{a}$ defined as

$$
P(\boldsymbol{a}) = \begin{bmatrix} \mathcal{R}_{\boldsymbol{a}}(0) & \mathcal{R}_{\boldsymbol{a}}(1) & \cdots & \mathcal{R}_{\boldsymbol{a}}(M-1) \\ \mathcal{R}_{\boldsymbol{a}}(1) & \mathcal{R}_{\boldsymbol{a}}(0) & \cdots & \mathcal{R}_{\boldsymbol{a}}(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{R}_{\boldsymbol{a}}(M-1) & \mathcal{R}_{\boldsymbol{a}}(M-2) & \cdots & \mathcal{R}_{\boldsymbol{a}}(0) \end{bmatrix}, \tag{6}
$$

where $\mathcal{R}_{\boldsymbol{a}}(\tau) := \sum_{i=1}^{N-\tau} a_i a_{i+\tau}$ denotes the un-normalized sample autocorrelation function of $\boldsymbol{a} \in \mathbb{R}^N$.

Theorem 1: Let $\boldsymbol{a} \in \mathbb{R}^N$ be fixed. Define two quantities $\rho(\boldsymbol{a})$ and $\mu(\boldsymbol{a})$ associated with the eigenvalues of the covariance matrix $P(\boldsymbol{a})$ as $\rho(\boldsymbol{a}) := \frac{\max_i \lambda_i}{\|\boldsymbol{a}\|_2^2}$ and $\mu(\boldsymbol{a}) := \frac{\sum_{i=1}^M \lambda_i^2}{M\|\boldsymbol{a}\|_2^4}$, where λ_i is the i-th eigenvalue of $P(\boldsymbol{a})$. Let $\boldsymbol{y} = X\boldsymbol{a}$, where X is a random compressive Toeplitz matrix with i.i.d. Gaussian entries having zero mean and unit variance. Noting that $\mathbf{E}\left[\|\boldsymbol{y}\|_2^2\right] = M\|\boldsymbol{a}\|_2^2$, then for any $\epsilon \in (0,1)$, the upper tail probability bound is

$$
\mathbf{P}\left\{ \|\boldsymbol{y}\|_2^2 - M\|\boldsymbol{a}\|_2^2 \geq \epsilon M \|\boldsymbol{a}\|_2^2 \right\} \leq e^{-\frac{\epsilon^2 M}{8\rho(\boldsymbol{a})}} \tag{7}
$$

and the lower tail probability bound is

$$
\mathbf{P}\left\{ \|\boldsymbol{y}\|_2^2 - M\|\boldsymbol{a}\|_2^2 \leq -\epsilon M \|\boldsymbol{a}\|_2^2 \right\} \leq e^{-\frac{\epsilon^2 M}{8\mu(\boldsymbol{a})}}. \tag{8}
$$

Theorem 1 provides CoM inequalities for *any* (not necessarily sparse) signal $\boldsymbol{a} \in \mathbb{R}^N$. The significance of these results comes from the fact that the tail probability bounds are functions of the signal $\boldsymbol{a}$, where the dependency is captured in the quantities $\rho(\boldsymbol{a})$ and $\mu(\boldsymbol{a})$. This is not the case when X is unstructured.

Indeed, allowing X to have $M \times N$ i.i.d. Gaussian entries with zero mean and unit variance (and thus, no Toeplitz structure) would result in the concentration inequality (see, e.g., [3])

$$\mathbf{P} \left\{ \left| \|\boldsymbol{y}\|_2^2 - M\|\boldsymbol{a}\|_2^2 \right| \geq \epsilon M\|\boldsymbol{a}\|_2^2 \right\} \leq 2e^{-\frac{\epsilon^2 M}{4}}. \tag{9}$$

Thus, comparing the bound in (9) with the ones in (7) and (8), one could conclude that achieving the same probability bound for Toeplitz matrices requires choosing M larger by a factor of $2\rho(\boldsymbol{a})$ or $2\mu(\boldsymbol{a})$. Typically, when using CoM inequalities such as (7) and (8), we must set M large enough so that both bounds are sufficiently small over all signals $\boldsymbol{a}$ belonging to some class of interest. For example, we are often interested in signals that have a *sparse* representation. Because we generally wish to keep M as small as possible, it is interesting to try to obtain an upper bound for the important quantities $\rho(\boldsymbol{a})$ and $\mu(\boldsymbol{a})$ over the class of signals of interest. It is easy to show that for all $\boldsymbol{a} \in \mathbb{R}^N$, $\mu(\boldsymbol{a}) \leq \rho(\boldsymbol{a})$. Thus, we limit our analysis to finding the sharpest upper bound for $\rho(\boldsymbol{a})$ when $\boldsymbol{a}$ is K-sparse. For the sake of generality, we allow the signal to be sparse in an arbitrary orthobasis.

Definition 1: A signal $\boldsymbol{a} \in \mathbb{R}^N$ is called K-*sparse in an orthobasis* $G \in \mathbb{R}^{N \times N}$ if it can be represented as $\boldsymbol{a} = G\boldsymbol{q}$, where $\boldsymbol{q} \in \mathbb{R}^N$ is K-sparse (a vector with $K < N$ non-zero entries).

We also introduce the notion of K-*sparse Fourier coherence* of the orthobasis G. This measures how strongly the columns of G are correlated with the length L Fourier basis, $F_L \in \mathbb{C}^{L \times L}$, which has entries $F_L(\ell, m) = \frac{1}{\sqrt{L}} w^{(\ell-1)(m-1)}$, where $w = e^{-\frac{2\pi j}{L}}$.

Definition 2: Given an orthobasis $G \in \mathbb{R}^{N \times N}$ and measurement length M, let $L = N + M - 1$. The K-*sparse Fourier coherence of* G, denoted $\nu_K(G)$, is defined as

$$\nu_K(G) := \max_{i,S} \|F_{1:N}^{i \to} G_S\|_2, \tag{10}$$

where $S \subseteq \{1, 2, \ldots, N\}$ is the *support set* and varies over all possible sets with cardinality $|S| = K$, $G_S \in \mathbb{R}^{N \times K}$ is a matrix containing the columns of $G \in \mathbb{R}^{N \times N}$ indexed by the support set S, and $F_{1:N}^{i \to} \in \mathbb{C}^N$ is a row vector containing the first N entries of the i-th row of the Fourier orthobasis $F_L \in \mathbb{C}^{L \times L}$. Observe that for a given orthobasis G, $\nu_K(G)$ depends on K.

Using the notion of Fourier coherence, we show in Section IV that for all vectors $\boldsymbol{a} \in \mathbb{R}^N$ that are K-sparse in an orthobasis $G \in \mathbb{R}^{N \times N}$,

$$\rho(\boldsymbol{a}) \leq L\nu_K^2(G), \tag{11}$$

where, as above, $L = N + M - 1$. This bound, however, appears to be highly pessimistic for most K-sparse signals. As a step towards better understanding the behavior of $\rho(\boldsymbol{a})$, we consider a random model for $\boldsymbol{a}$. In particular, we consider a fixed K-sparse support set, and on this set we suppose the K

non-zero entries of the coefficient vector $\boldsymbol{q}$ are drawn from a random distribution. Based on this model, we derive an upper bound for $\mathbf{E}\left[\rho\left(\boldsymbol{a}\right)\right]$.

Theorem 2: (*Upper Bound on* $\mathbf{E}\left[\rho\left(\boldsymbol{a}\right)\right]$) Let $\boldsymbol{q} \in \mathbb{R}^N$ be a random K-sparse vector whose K non-zero entries (on an arbitrary support S) are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \frac{1}{K})$. Select the measurement length M, which corresponds to the dimension of $P(\boldsymbol{a})$, and set $L = N + M - 1$. Let $\boldsymbol{a} = G\boldsymbol{q}$ where $G \in \mathbb{R}^{N \times N}$ is an orthobasis. Then

$$\mathbf{E}\left[\rho\left(\boldsymbol{a}\right)\right] \leq \frac{8L\nu_K^2\left(G\right)}{K}\left(\log 2L + 2\right). \tag{12}$$

The K-sparse Fourier coherence $\nu_K\left(G\right)$ and consequently the bounds (11) and (12) can be explicitly evaluated for some specific orthobases G. For example, letting $G = I_N$ (the $N \times N$ identity matrix), we can consider signals that are sparse in the time domain. With this choice of G, one can show that $\nu_K\left(I_N\right) = \sqrt{\frac{K}{L}}$. As another example, we can consider signals that are sparse in the frequency domain. To do this, we set G equal to a real-valued version of the Fourier orthobasis. Without loss of generality, suppose N is even. The real Fourier orthobasis, denoted R_N, is constructed as follows. The first column of R_N equals the first column of F_N. Then $R_{\{2,\ldots,\frac{N}{2}\}} = \text{Real}\left(\sqrt{2}F_{\{2,\ldots,\frac{N}{2}\}}\right)$ and $R_{\{\frac{N}{2}+1,\ldots,N-1\}} = \text{Imaginary}\left(\sqrt{2}F_{\{2,\ldots,\frac{N}{2}\}}\right)$. The last column of R_N is equal to the $(\frac{N}{2}+1)$-th column of F_N. Similar steps can be taken to construct a real Fourier orthobasis when N is odd. With this choice of G, one can show that $\nu_K\left(R_N\right) \leq \sqrt{\frac{N}{L}}$. Using these upper bounds on the Fourier coherence, we have the following deterministic bounds on $\rho\left(\boldsymbol{a}\right)$ in the time and frequency domains:

$$\rho\left(\boldsymbol{a}\right) \leq K \quad \text{(time domain sparsity)} \quad \text{and} \tag{13}$$

$$\rho\left(\boldsymbol{a}\right) \leq N \quad \text{(frequency domain sparsity)}. \tag{14}$$

We also obtain the following bounds on the expected value of $\rho\left(\boldsymbol{a}\right)$ under the random signal model:

$$\mathbf{E}\left[\rho\left(\boldsymbol{a}\right)\right] \leq 8\left(\log 2L + 2\right) \quad \text{(time domain sparsity)} \quad \text{and} \tag{15}$$

$$\mathbf{E}\left[\rho\left(\boldsymbol{a}\right)\right] \leq \frac{8N}{K}\left(\log 2L + 2\right) \quad \text{(frequency domain sparsity)}. \tag{16}$$

We offer a brief interpretation and analysis of these bounds in this paragraph and several examples that follow. First, because $K \leq N$, the deterministic and expectation bounds on $\rho\left(\boldsymbol{a}\right)$ are smaller for signals that are sparse in the time domain than for signals that are sparse in the frequency domain. The simulations described in Examples 2 and 3 below confirm that, on average, $\rho\left(\boldsymbol{a}\right)$ does indeed tend to be smaller under the model of time domain sparsity. Second, these bounds exhibit varying dependencies on

the sparsity level K: (13) increases with K and (16) decreases with K, while (14) and (15) are agnostic to K. The simulation described in Example 2 below confirms that, on average, $\rho(\boldsymbol{a})$ increases with K for signals that are sparse in the time domain but decreases with K for signals that are sparse in the frequency domain. This actually reveals a looseness in (15); however, in Section IV-C, we conjecture a sparsity-dependent expectation bound that closely matches the empirical results for signals that are sparse in the time domain. Third, under both models of sparsity, and assuming $8(\log 2L + 2) \ll K$ for signals of practical interest, the expectation bounds on $\rho(\boldsymbol{a})$ are qualitatively lower than the deterministic bounds. This raises the question of whether the deterministic bounds are sharp. We confirm that this is the case in Example 4 below.

Example 1: (Illustrating the signal-dependency of the left-hand side of CoM inequalities (7) and (8)) In this example, we illustrate that the CoM behavior for randomized Toeplitz matrices is indeed signal-dependent. We consider inequality (7) while a similar analysis can be made for (8). We consider two particular K-sparse ($K = 64$) signals, $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ both in $\mathbb{R}^N$ ($N = 1024$) where the K non-zero entries of $\boldsymbol{a}_1$ have equal values and occur in the first K entries of the vector ($\rho(\boldsymbol{a}_1) = 63.26$), while the K non-zero entries of $\boldsymbol{a}_2$ appear in a randomly-selected locations with random signs and values ($\rho(\boldsymbol{a}_2) = 5.47$). Both $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are normalized. For a fixed $M = 512$, we measure each of these signals with 1000 i.i.d. Gaussian $M \times N$ Toeplitz matrices. Figure 1 depicts the numerically determined rate of occurrence of the event $\|\boldsymbol{y}\|_2^2 - M\|\boldsymbol{a}\|_2^2 \geq \epsilon M \|\boldsymbol{a}\|_2^2$ over 1000 trials versus $\epsilon \in (0,1)$. For comparison, the derived analytical bound in (7) (for Toeplitz X) as well as the bound in (9) (for unstructured X) is depicted. As can be seen the two signals have different concentrations. In particular, $\boldsymbol{a}_1$ (Fig. 1(a)) has worse concentration compared to $\boldsymbol{a}_2$ (Fig. 1(b)) when measured by Toeplitz matrices. This signal-dependency of the concentration does not happen when these signals are measured by unstructured Gaussian random matrices. Moreover, the derived signal-dependent bound (7) successfully upper bounds the numerical event rate of occurrence for each signal while the bound (9) for unstructured X fails to do so. Also observe that the analytical bound $e^{-\frac{M\epsilon^2}{8\rho(\boldsymbol{a}_2)}}$ in Fig. 1(b) can not bound the numerical event rate of occurrence for $\boldsymbol{a}_1$ as depicted in Fig. 1(a).

Example 2: (Varying K and comparing the time and frequency domains) In this experiment, we fix M and N. For each value of K and each sparse basis $G = I_N$ and $G = R_N$, we construct 1000 random sparse vectors $\boldsymbol{q} \in \mathbb{R}^N$ with random support and having K non-zero entries drawn from a Gaussian distribution with mean zero and variance $\frac{1}{K}$. For each vector, we compute $\boldsymbol{a} = G\boldsymbol{q}$, and we then let $\bar{\rho}(\boldsymbol{a})$ denote the sample mean of $\rho(\boldsymbol{a})$ across these 1000 signals. The results, as a function of K, are plotted in Fig. 2(a). As anticipated, signals that are sparse in the frequency domain have a larger value of $\bar{\rho}(\boldsymbol{a})$ than
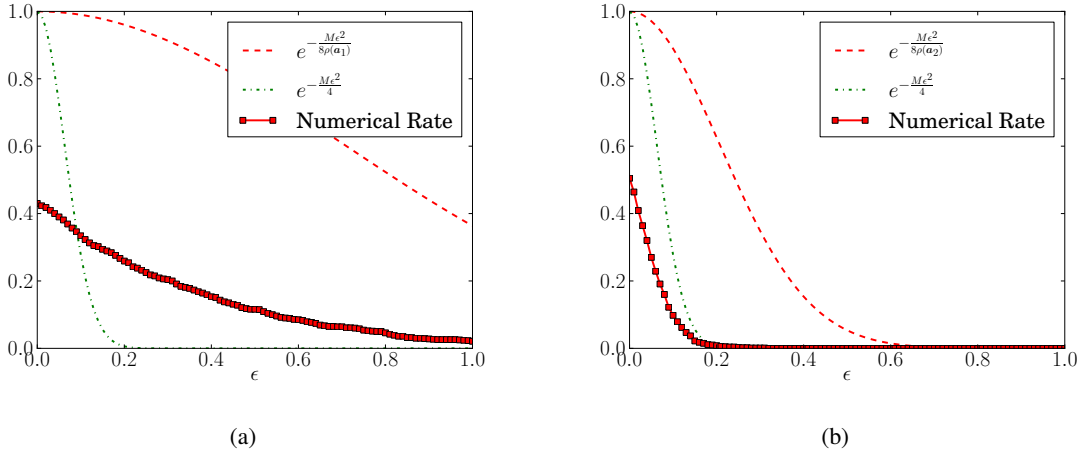
Fig. 1: *Illustrating the signal-dependency of the left-hand side of CoM inequality* (7). *With fixed* $N = 1024$, $K = 64$ *and* $M = 512$, *we consider two particular* K-*sparse signals,* $\boldsymbol{a_1}$, *and* $\boldsymbol{a_2}$. *Both* $\boldsymbol{a_1}$ *and* $\boldsymbol{a_2}$ *are normalized. We measure each of these signals with* 1000 *i.i.d. Gaussian* $M \times N$ *Toeplitz matrices. (a) The* K *non-zero entries of* $\boldsymbol{a_1}$ *have equal values and occur in the first* K *entries of the vector. (b) The* K *non-zero entries of* $\boldsymbol{a_2}$ *appear in randomly-selected locations with random signs and values. The two signals have different concentrations that can be upper bounded by the signal-dependent right-hand side of* (7).

signals that are sparse in the time domain. Moreover, $\bar{\rho}(\boldsymbol{a})$ decreases with K in the frequency domain but increases with K in the time domain. Overall, the empirical behavior of $\bar{\rho}(\boldsymbol{a})$ is mostly consistent with the bounds (15) and (16), although our constants may be larger than necessary.

Example 3: (Varying M and comparing the time and frequency domains) This experiment is identical to the one in Example 2, except that we vary M while keeping K and N fixed. The results are plotted in Fig. 2(b). Once again, signals that are sparse in the frequency domain have a larger value of $\bar{\rho}(\boldsymbol{a})$ than signals that are sparse in the time domain. Moreover, in both cases $\bar{\rho}(\boldsymbol{a})$ appears to increase logarithmically with M as predicted by the bounds in (15) and (16), although our constants may be larger than necessary.

Example 4: (Confirming the tightness of the deterministic bounds) We fix N, consider a vector $\boldsymbol{a} \in \mathbb{R}^N$ that is K-sparse in the time domain, and suppose the K non-zero entries of $\boldsymbol{a}$ take equal values and occur in the first K entries of the vector. For such a vector one can derive a lower bound on $\rho(\boldsymbol{a})$ by embedding $P(\boldsymbol{a})$ inside a circulant matrix, applying the Cauchy Interlacing Theorem [23] (we describe these steps more fully in Section IV-A1), and then performing further computations that we omit for the
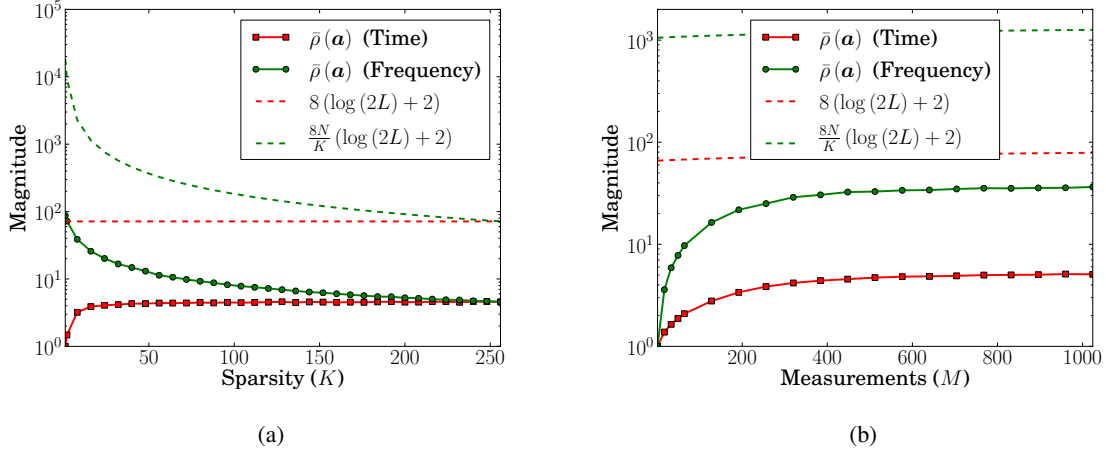
Fig. 2: *Sample mean of $\rho(\boldsymbol{a})$ in the time and frequency domains versus the expectation bounds (15) and (16), where $L = N + M - 1$. The sample mean $\bar{\rho}(\boldsymbol{a})$ is calculated by taking the mean over 1000 constructed signals $\boldsymbol{a}$. A logarithmic scale is used for the vertical axis. (a) Example 2: Varying K with fixed $M = N = 256$. (b) Example 3: Varying M with fixed $N = 512$ and $K = 32$.*

sake of space. With these steps, one concludes that for this specific vector $\boldsymbol{a} \in \mathbb{R}^N$,

$$\rho(\boldsymbol{a}) \geq K \left(1 - \frac{\pi^2}{24} \left(\frac{K^2}{M + K - 1} \right)^2 \right)^2. \tag{17}$$

When $M \gg K^2$, the right-hand side of (17) approaches K. This confirms that (13) is sharp for large M.

In the remainder of the paper, the proofs of Theorem 1 (Section III) and Theorem 2 (Section IV) are presented, followed by additional discussions concerning the relevance of the main results. Our results have important consequences in the analysis of high-dimensional dynamical systems. We expound on this fact by exploring a CBD problem in Section V.

III. PROOF OF THEOREM 1

The proofs of the upper and lower bounds of Theorem 1 are given separately in Lemmas 2 and 3 below. The proofs utilize Markov's inequality along with a suitable bound on the moment generating function of $\|\boldsymbol{y}\|_2^2$, given in Lemma 1. Observe that for a fixed vector $\boldsymbol{a} \in \mathbb{R}^N$ and a random Gaussian Toeplitz matrix $X \in \mathbb{R}^{M \times N}$, the vector $\boldsymbol{y} = X\boldsymbol{a} \in \mathbb{R}^M$ will be a Gaussian random vector with zero mean and $M \times M$ covariance matrix $P(\boldsymbol{a})$ given in (6).

Lemma 1: If $\boldsymbol{y} \in \mathbb{R}^M$ is a zero mean Gaussian random vector with covariance matrix P, then

$$\mathbf{E}\left[e^{t\boldsymbol{y}^T\boldsymbol{y}}\right] = \frac{1}{\sqrt{\det\left(I_M - 2tP\right)}} \tag{18}$$

holds for all $t \in (-\infty, \frac{1}{2\lambda_{\max}(P)})$.

Proof

$$\mathbf{E}\left[e^{t\boldsymbol{y}^T\boldsymbol{y}}\right] = \int \frac{1}{(2\pi)^{\frac{M}{2}} \det^{\frac{1}{2}}(P)} e^{t\boldsymbol{y}^T\boldsymbol{y}} e^{-\frac{1}{2}\boldsymbol{y}^T P^{-1}\boldsymbol{y}} d\boldsymbol{y}$$

$$= \int \frac{1}{(2\pi)^{\frac{M}{2}} \det^{\frac{1}{2}}(P)} e^{-\frac{1}{2}\boldsymbol{y}^T(P^{-1}-2tI_M)\boldsymbol{y}} d\boldsymbol{y} = \frac{\det^{\frac{1}{2}}\left(\left(P^{-1}-2tI_M\right)^{-1}\right)}{\det^{\frac{1}{2}}(P)}$$

$$= \frac{1}{\left(\det\left(P^{-1}-2tI_M\right)\det P\right)^{\frac{1}{2}}} = \frac{1}{\sqrt{\det\left(I_M - 2tP\right)}}.$$

∎

Observe that as a special case of Lemma 1, if $y \in \mathbb{R}$ is a scalar Gaussian random variable of unit variance, then we obtain the well known result of $\mathbf{E}\left[e^{ty^2}\right] = \frac{1}{\sqrt{1-2t}}$, for $t \in (-\infty, \frac{1}{2})$. Based on Lemma 1, we use Chernoff's bounding method [24] for computing the upper and lower tail probability bounds. In particular, we are interested in finding bounds for the tail probabilities

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 \geq M\|\boldsymbol{a}\|_2^2(1+\epsilon)\right\} \tag{19a}$$

and

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 \leq M\|\boldsymbol{a}\|_2^2(1-\epsilon)\right\}. \tag{19b}$$

Observe that in (19a) and (19b) concentration behavior is sought around $\mathbf{E}\left[\|\boldsymbol{y}\|_2^2\right] = M\|\boldsymbol{a}\|_2^2$. For a random variable z, and all $t > 0$,

$$\mathbf{P}\left\{z \geq \epsilon\right\} = \mathbf{P}\left\{e^{tz} \geq e^{t\epsilon}\right\} \leq \frac{\mathbf{E}\left[e^{tz}\right]}{e^{t\epsilon}} \tag{20}$$

(see, e.g., [24]). Applying (20) to (19a), for example, and then applying Lemma 1 yields

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 > M\|\boldsymbol{a}\|_2^2(1+\epsilon)\right\} \leq \frac{\mathbf{E}\left[e^{t\boldsymbol{y}^T\boldsymbol{y}}\right]}{e^{M\|\boldsymbol{a}\|_2^2(1+\epsilon)t}} = \left(\det\left(I_M - 2tP\right)\right)^{-\frac{1}{2}} e^{-M\|\boldsymbol{a}\|_2^2(1+\epsilon)t}. \tag{21}$$

In (21), $t \in (-\infty, \frac{1}{2(\max_i \lambda_i)})$ is a free variable which—as in a Chernoff bound—can be varied to make the right-hand side as small as possible. Though not necessarily optimal, we propose to use $t = \frac{\epsilon}{2(1+\epsilon)f(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}$, where f is a function of $\boldsymbol{a}$ that we specify below. We state the upper tail probability bound in Lemma 2 and the lower tail probability bound in Lemma 3.

Lemma 2: Let $\boldsymbol{a} \in \mathbb{R}^N$ be fixed, let $P = P(\boldsymbol{a})$ be as given in (6), and let $\boldsymbol{y} \in \mathbb{R}^M$ be a zero mean Gaussian random vector with covariance matrix P. Then, for any $\epsilon \in (0, 1)$,

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 \geq M\|\boldsymbol{a}\|_2^2(1+\epsilon)\right\} \leq e^{-\frac{\epsilon^2 M}{8\rho(\boldsymbol{a})}}. \tag{22}$$

Proof Choosing t as

$$t = \frac{\epsilon}{2(1+\epsilon)\rho(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}$$

and noting that $t \in (-\infty, \frac{1}{2\max_i \lambda_i})$, the right-hand side of (21) can be written as

$$\left(\left(\det\left(I_M - \frac{\epsilon}{(1+\epsilon)}\frac{P}{\rho(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right)\right)^{-\frac{1}{M}} e^{-\frac{\epsilon}{\rho(\boldsymbol{a})}}\right)^{\frac{M}{2}}. \tag{23}$$

This expression can be simplified. Note that

$$\det\left(I_M - \frac{\epsilon}{(1+\epsilon)}\frac{P}{\rho(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right) = \prod_{i=1}^{M}\left(1 - \frac{\epsilon}{(1+\epsilon)}\frac{\lambda_i}{\rho(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right)$$

$$= e^{\sum_{i=1}^{M}\log\left(1 - \frac{\epsilon}{(1+\epsilon)}\frac{\lambda_i}{\rho(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right)}.$$

Using the facts that $\log(1 - c_1 c_2) \geq c_2 \log(1 - c_1)$ for any $c_1, c_2 \in [0, 1]$ and $\mathrm{Tr}(P) = M\|\boldsymbol{a}\|_2^2$, we have

$$e^{\sum_{i=1}^{M}\log\left(1 - \frac{\epsilon}{(1+\epsilon)}\frac{\lambda_i}{\rho(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right)} \geq e^{\sum_{i=1}^{M}\frac{\lambda_i}{\rho(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\log\left(1 - \frac{\epsilon}{1+\epsilon}\right)}$$

$$= e^{\frac{\mathrm{Tr}(P)}{\rho(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\log\left(\frac{1}{1+\epsilon}\right)} = e^{\frac{M}{\rho(\boldsymbol{a})}\log\left(\frac{1}{1+\epsilon}\right)} = \left(\frac{1}{1+\epsilon}\right)^{\frac{M}{\rho(\boldsymbol{a})}}. \tag{24}$$

Combining (21), (23), and (24) gives us

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 > M\|\boldsymbol{a}\|_2^2(1+\epsilon)\right\} \leq \left(\left(\frac{1}{1+\epsilon}\right)^{-\frac{1}{\rho(\boldsymbol{a})}} e^{-\frac{\epsilon}{\rho(\boldsymbol{a})}}\right)^{\frac{M}{2}} = \left((1+\epsilon)e^{-\epsilon}\right)^{\frac{M}{2\rho(\boldsymbol{a})}}.$$

The final bound comes by noting that $(1+\epsilon)e^{-\epsilon} \leq e^{-\epsilon^2/4}$. ∎

Lemma 3: Using the same assumptions as in Lemma 2, for any $\epsilon \in (0, 1)$,

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 \leq M\|\boldsymbol{a}\|_2^2(1-\epsilon)\right\} \leq e^{-\frac{\epsilon^2 M}{8\mu(\boldsymbol{a})}}.$$

Proof Applying Markov's inequality to (19b), we obtain

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 \leq M\|\boldsymbol{a}\|_2^2(1-\epsilon)\right\} = \mathbf{P}\left\{-\|\boldsymbol{y}\|_2^2 \geq -M\|\boldsymbol{a}\|_2^2(1-\epsilon)\right\} \leq \frac{\mathbf{E}\left[e^{-t\boldsymbol{y}^T\boldsymbol{y}}\right]}{e^{-M\|\boldsymbol{a}\|_2^2(1-\epsilon)t}}. \tag{25}$$

Using Lemma 1, this implies that

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 \le M\|\boldsymbol{a}\|_2^2(1-\epsilon)\right\} \le \left((\det(I_M + 2tP))^{-\frac{1}{M}} e^{2\|\boldsymbol{a}\|_2^2(1-\epsilon)t}\right)^{\frac{M}{2}}. \tag{26}$$

In this case, we choose

$$t = \frac{\epsilon}{2(1+\epsilon)\mu(\boldsymbol{a})\|\boldsymbol{a}\|_2^2},$$

and note that $t > 0$. Plugging t into (26) and following similar steps as for the upper tail bound, we get

$$\det(I_M + 2tP) = \prod_{i=1}^{M}\left(1 + \frac{\epsilon}{(1+\epsilon)}\frac{\lambda_i}{\mu(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right) = e^{\sum_{i=1}^{M}\log\left(1 + \frac{\epsilon}{(1+\epsilon)}\frac{\lambda_i}{\mu(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right)}. \tag{27}$$

Since $\log(1+c) \ge c - \frac{c^2}{2}$ for $c > 0$,

$$\sum_{i=1}^{M}\log\left(1 + \frac{\epsilon}{(1+\epsilon)}\frac{\lambda_i}{\mu(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right) \ge \sum_{i=1}^{M}\left(\frac{\epsilon}{(1+\epsilon)}\frac{\lambda_i}{\mu(\boldsymbol{a})\|\boldsymbol{a}\|_2^2} - \frac{1}{2}\left(\frac{\epsilon}{(1+\epsilon)}\frac{\lambda_i}{\mu(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right)^2\right)$$

$$= \frac{\epsilon}{(1+\epsilon)}\frac{\sum_{i=1}^{M}\lambda_i}{\mu(\boldsymbol{a})\|\boldsymbol{a}\|_2^2} - \frac{1}{2}\left(\frac{\epsilon}{(1+\epsilon)\mu(\boldsymbol{a})\|\boldsymbol{a}\|_2^2}\right)^2 \sum_{i=1}^{M}\lambda_i^2$$

$$= \frac{\epsilon}{(1+\epsilon)}\frac{M}{\mu(\boldsymbol{a})} - \frac{1}{2}\left(\frac{\epsilon}{1+\epsilon}\right)^2\frac{M}{\mu(\boldsymbol{a})} = \frac{M}{\mu(\boldsymbol{a})}\left(\frac{\epsilon^2 + 2\epsilon}{2(1+\epsilon)^2}\right). \tag{28}$$

Combining (27) and (28) gives the bound

$$\det(I_M + 2tP) \ge e^{\frac{M}{\mu(\boldsymbol{a})}\left(\frac{\epsilon^2+2\epsilon}{2(1+\epsilon)^2}\right)} = \left(e^{\frac{\epsilon^2+2\epsilon}{2(1+\epsilon)^2}}\right)^{\frac{M}{\mu(\boldsymbol{a})}}. \tag{29}$$

By substituting (29) into (26), we obtain

$$\mathbf{P}\left\{\|\boldsymbol{y}\|_2^2 \le M\|\boldsymbol{a}\|_2^2(1-\epsilon)\right\} \le \left(e^{\frac{-\epsilon^2-2\epsilon}{2(1+\epsilon)^2}} e^{\frac{\epsilon(1-\epsilon)}{1+\epsilon}}\right)^{\frac{M}{2\mu(\boldsymbol{a})}} = \left(e^{\frac{-2\epsilon^3-\epsilon^2}{2(1+\epsilon)^2}}\right)^{\frac{M}{2\mu(\boldsymbol{a})}}.$$

The final bound comes by noting that $e^{\frac{-2\epsilon^3-\epsilon^2}{2(1+\epsilon)^2}} \le e^{-\epsilon^2/4}$. ∎

IV. PROOF AND DISCUSSION OF THEOREM 2

A. Proof of Theorem 2

1) Circulant Embedding: The covariance matrix $P(\boldsymbol{a})$ described in (6) is an $M \times M$ symmetric Toeplitz matrix which can be decomposed as $P(\boldsymbol{a}) = A^T A$, where A is an $(N + M - 1) \times M$ Toeplitz matrix (as shown in Fig. 3) and A^T is the transpose of A. In order to derive an upper bound on the maximum eigenvalue of $P(\boldsymbol{a})$, we embed the matrix A inside its $(N + M - 1) \times (N + M - 1)$ circulant counterpart A_c where each column of A_c is a cyclic downward shifted version of the previous column. Thus, A_c is uniquely determined by its first column, which we denote by

$$\widetilde{\boldsymbol{a}} = [\underbrace{a_1 \cdots a_N}_{\boldsymbol{a}^T}\ \underbrace{0 \cdots 0}_{(M-1)\ \text{zeros}}]^T \in \mathbb{R}^L,$$

$$A = \begin{bmatrix} a_1 & 0 & \dots & & & 0 \\ a_2 & \ddots & & & (0) & \\ \vdots & \ddots & a_1 & & & \vdots \\ a_N & & a_2 & \ddots & & 0 \\ 0 & \ddots & \vdots & \ddots & & a_1 \\ \vdots & & a_N & & & a_2 \\ & (0) & & & \ddots & \vdots \\ 0 & & \dots & 0 & & a_N \end{bmatrix} \qquad A_c = \begin{bmatrix} a_1 & 0 & \dots & & 0 & a_N & \dots & a_2 \\ a_2 & \ddots & \ddots & & & & \ddots & \ddots & \vdots \\ \vdots & \ddots & a_1 & & (0) & & & \ddots & a_N \\ a_N & & a_2 & \ddots & & & & & 0 \\ 0 & \ddots & \vdots & \ddots & & a_1 & & & \\ \vdots & & a_N & & & a_2 & \ddots & \ddots & \vdots \\ & (0) & & & \ddots & \vdots & \ddots & \ddots & 0 \\ 0 & & \dots & 0 & & a_N & \dots & a_2 & a_1 \end{bmatrix}$$

Fig. 3: *Toeplitz matrix $A \in \mathbb{R}^{L \times M}$ and its circulant counterpart $A_c \in \mathbb{R}^{L \times L}$ where $L = N + M - 1$.*

where $L = N + M - 1$. Observe that the circulant matrix $A_c \in \mathbb{R}^{L \times L}$ contains the Toeplitz matrix $A \in \mathbb{R}^{L \times M}$ in its first M columns. Because of this embedding, the Cauchy Interlacing Theorem [23] implies that $\max_{i \leq M} \lambda_i(A^T A) \leq \max_{i \leq L} \lambda_i(A_c^T A_c)$. Therefore, we have

$$\rho(\boldsymbol{a}) = \frac{\max_i \lambda_i(P(\boldsymbol{a}))}{\|\boldsymbol{a}\|_2^2} = \frac{\max_i \lambda_i(A^T A)}{\|\boldsymbol{a}\|_2^2} \leq \frac{\max_i \lambda_i(A_c^T A_c)}{\|\boldsymbol{a}\|_2^2} = \frac{\max_i |\lambda_i(A_c^T)|^2}{\|\boldsymbol{a}\|_2^2} =: \rho_c(\boldsymbol{a}). \tag{30}$$

Thus, an upper bound for $\rho(\boldsymbol{a})$ can be achieved by bounding the maximum absolute eigenvalue of A_c^T. Since A_c^T is circulant, its eigenvalues are given by the un-normalized length-L Discrete Fourier Transform (DFT) of the first row of A_c^T (the first column of A_c). Specifically, for $i = 1, 2, \dots, L$,

$$\lambda_i(A_c^T) = \sum_{k=1}^{L} \widetilde{a}_k e^{-\frac{2\pi j}{L}(i-1)(k-1)} = \sum_{k=1}^{N} a_k e^{-\frac{2\pi j}{L}(i-1)(k-1)}. \tag{31}$$

Recall that $F_L \in \mathbb{C}^{L \times L}$ is the Fourier orthobasis with entries $F_L(\ell, m) = \frac{1}{\sqrt{L}} w^{(\ell-1)(m-1)}$ where $w = e^{-\frac{2\pi j}{L}}$, and let $F_L^{i \to} \in \mathbb{C}^L$ be the i-th row of F_L. Using matrix-vector notation, (31) can be written as

$$\lambda_i(A_c^T) = \sqrt{L} F_L^{i \to} \widetilde{\boldsymbol{a}} = \sqrt{L} F_{1:N}^{i \to} \boldsymbol{a} = \sqrt{L} F_{1:N}^{i \to} G \boldsymbol{q} = \sqrt{L} F_{1:N}^{i \to} G_S \boldsymbol{q}_S, \tag{32}$$

where $F_{1:N}^{i \to} \in \mathbb{C}^N$ is a row vector containing the first N entries of $F_L^{i \to}$, $\boldsymbol{q}_S \in \mathbb{R}^K$ is the part of $\boldsymbol{q} \in \mathbb{R}^N$ restricted to the support S (the location of the non-zero entries of $\boldsymbol{q}$) with cardinality $|S| = K$, and $G_S \in \mathbb{R}^{N \times K}$ contains the columns of $G \in \mathbb{R}^{N \times N}$ indexed by the support S.

2) Deterministic Bound: We can bound $\rho(\boldsymbol{a})$ over all sparse $\boldsymbol{a}$ using the Cauchy-Schwarz inequality. From (32), it follows for any $i \in \{1, 2, \dots, L\}$ that

$$|\lambda_i(A_c^T)| = |\sqrt{L} F_{1:N}^{i \to} G_S \boldsymbol{q}_S| \leq \sqrt{L} \|F_{1:N}^{i \to} G_S\|_2 \|\boldsymbol{q}_S\|_2 = \sqrt{L} \|F_{1:N}^{i \to} G_S\|_2 \|\boldsymbol{a}\|_2. \tag{33}$$

By combining Definition 2, (30), and (33), we arrive at the deterministic bound (11). This bound appears to be highly pessimistic for *most* sparse vectors $\boldsymbol{a}$. In other words, although in Example 4 we illustrate that for a specific signal $\boldsymbol{a}$, the deterministic bound (11) is tight when $M \gg K$, we observe that for many other classes of sparse signals $\boldsymbol{a}$, the bound is pessimistic. In particular, if a random model is imposed on the non-zero entries of $\boldsymbol{a}$, an upper bound on the typical value of $\rho(\boldsymbol{a})$ derived in (15) scales logarithmically in the ambient dimension L which is qualitatively smaller than K. We show this analysis in the proof of Theorem 2. In order to make this proof self-contained, we first list some results that we will draw from.

3) Supporting Results: We utilize the following propositions.

Lemma 4: [19] Let z be any random variable. Then

$$\mathbf{E}\left[|z|\right] = \int_0^\infty \mathbf{P}\left\{|z| \geq x\right\} dx. \tag{34}$$

Lemma 5: Let z_1 and z_2 be positive random variables. Then for any U,

$$\mathbf{P}\left\{z_1 + z_2 \geq U\right\} \leq \mathbf{P}\left\{z_1 \geq \frac{U}{2}\right\} + \mathbf{P}\left\{z_2 \geq \frac{U}{2}\right\}, \tag{35}$$

and for any U_1 and U_2,

$$\mathbf{P}\left\{\frac{z_1}{z_2} \geq \frac{U_1}{U_2}\right\} \leq \mathbf{P}\left\{z_1 \geq U_1\right\} + \mathbf{P}\left\{z_2 \leq U_2\right\}. \tag{36}$$

Proof See Appendix A. ∎

Proposition 1: [3] (*Concentration Inequality for Sums of Squared Gaussian Random Variables*) Let $\boldsymbol{q} \in \mathbb{R}^N$ be a random K-sparse vector whose K non-zero entries (on an arbitrary support S) are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \sigma^2)$. Then for any $\epsilon > 0$,

$$\mathbf{P}\left\{\|\boldsymbol{q}\|_2^2 \leq K\sigma^2 (1 - \epsilon)\right\} \leq e^{-\frac{K\epsilon^2}{4}}.$$

Proposition 2: (*Hoeffding's Inequality for Complex-Valued Gaussian Sums*) Let $\boldsymbol{b} \in \mathbb{C}^N$ be fixed, and let $\boldsymbol{\epsilon} \in \mathbb{R}^N$ be a random vector whose N entries are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \sigma^2)$. Then, for any $u > 0$,

$$\mathbf{P}\left\{\left|\sum_{i=1}^N \epsilon_i b_i\right| \geq u\right\} \leq 2e^{-\frac{u^2}{4\sigma^2 \|\boldsymbol{b}\|_2^2}}.$$

Proof See Appendix B. ∎

In order prove Theorem 2, we also require a tail probability bound for the eigenvalues of A_c^T.

Proposition 3: Let $\boldsymbol{q} \in \mathbb{R}^N$ be a random K-sparse vector whose K non-zero entries (on an arbitrary support S) are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \frac{1}{K})$. Let $\boldsymbol{a} = G\boldsymbol{q}$ where $G \in \mathbb{R}^{N \times N}$ is an orthobasis, and let A_c be an $L \times L$ circulant matrix, where the first N entries of the first column of A_c are given by $\boldsymbol{a}$. Then for any $u > 0$, and for $i = 1, 2, \ldots, L$,

$$\mathbf{P}\left\{|\lambda_i(A_c)| \geq u\right\} \leq 2e^{-\frac{u^2 K}{4L\nu_K^2(G)}}. \tag{37}$$

Proof Define the row vector $\boldsymbol{b} = \sqrt{L} F_{1:N}^{i\to} G_S \in \mathbb{C}^K$. From (32) and the Cauchy-Schwarz inequality, it follows that $|\lambda_i(A_c)| = |\lambda_i(A_c^T)| = |\sqrt{L} F_{1:N}^{i\to} G_S \boldsymbol{q}_S| = |\sum_{i=1}^K \epsilon_i b_i|$, where $\epsilon_i = (\boldsymbol{q}_S)_i$. From Definition 2, we have $\|\boldsymbol{b}\|_2 \leq \sqrt{L}\nu_K(G)$. The tail probability bound (37) follows from applying Proposition 2. ∎

4) Completing the Proof of Theorem 2: From (30), we have

$$\mathbf{E}[\rho(\boldsymbol{a})] \leq \mathbf{E}[\rho_c(\boldsymbol{a})] = \mathbf{E}\left[\frac{\max_i |\lambda_i(A_c^T)|^2}{\|\boldsymbol{a}\|_2^2}\right] = \int_0^\infty \mathbf{P}\left\{\frac{\max_i |\lambda_i(A_c^T)|^2}{\|\boldsymbol{a}\|_2^2} \geq x\right\} dx$$

$$= \int_0^{L\nu_G^2} \mathbf{P}\left\{\frac{\max_i |\lambda_i(A_c^T)|^2}{\|\boldsymbol{a}\|_2^2} \geq x\right\} dx,$$

where the last equality comes from the deterministic upper bound $|\lambda_i(A_c^T)| \leq \sqrt{L}\|F_{1:N}^{i\to} G_S\|_2 \|\boldsymbol{a}\|_2 \leq \sqrt{L}\nu_K(G)\|\boldsymbol{a}\|_2$. Using a union bound, for any $t > 0$ we have

$$\int_0^{L\nu_K^2(G)} \mathbf{P}\left\{\frac{\max_i |\lambda_i(A_c^T)|^2}{\|\boldsymbol{a}\|_2^2} \geq x\right\} dx = \int_0^{L\nu_K^2(G)} \mathbf{P}\left\{\frac{\max_i |\lambda_i(A_c^T)|^2}{\|\boldsymbol{a}\|_2^2} \geq \frac{tx}{t}\right\} dx$$

$$\leq \int_0^{L\nu_K^2(G)} \mathbf{P}\left\{\max_i |\lambda_i(A_c^T)|^2 \geq tx\right\} dx$$

$$+ \int_0^{L\nu_K^2(G)} \mathbf{P}\left\{\|\boldsymbol{a}\|_2^2 \leq t\right\} dx. \tag{38}$$

The first term in the right hand side of (38) can be bounded as follows. For every $\delta \geq 0$, by partitioning the range of integration [19], [25], we obtain

$$
\int_0^{L\nu_K^2(G)} \mathbf{P}\left\{\max_i |\lambda_i(A_c^T)|^2 \geq tx\right\} dx \leq \int_0^\infty \mathbf{P}\left\{\max_i |\lambda_i(A_c^T)|^2 \geq tx\right\} dx
$$
$$
\leq \delta + \int_\delta^\infty \mathbf{P}\left\{\max_i |\lambda_i(A_c)|^2 \geq tx\right\} dx
$$
$$
\leq \delta + \int_\delta^\infty \sum_{i=1}^L \mathbf{P}\left\{|\lambda_i(A_c)|^2 \geq tx\right\} dx
$$
$$
\leq \delta + \int_\delta^\infty \sum_{i=1}^L 2e^{-\frac{Ktx}{4L\nu_K^2(G)}} dx
$$
$$
= \delta + 2L \int_\delta^\infty e^{-\frac{Ktx}{4L\nu_K^2(G)}} dx
$$
$$
= \delta + \frac{8L^2\nu_K^2(G)}{Kt} e^{-\frac{Kt\delta}{4L\nu_K^2(G)}},
$$

where we used Proposition 3 in the last inequality. The second term in (38) can be bounded using the concentration inequality of Proposition 1. We have for $0 < t \leq 1$, $\mathbf{P}\left\{\|\boldsymbol{a}\|_2^2 \leq t\right\} \leq e^{-\frac{K(1-t)^2}{4}}$. Putting together the bounds for the two terms of inequality (38), we have

$$
\mathbf{E}\left[\rho(\boldsymbol{a})\right] \leq \mathbf{E}\left[\rho_c(\boldsymbol{a})\right] \leq \delta + \frac{8L^2\nu_K^2(G)}{Kt} e^{-\frac{Kt\delta}{4L\nu_K^2(G)}} + L\nu_K^2(G) e^{-\frac{K(1-t)^2}{4}}. \tag{39}
$$

Now we pick δ to minimize the upper bound in (39). Using the minimizer $\delta^\star = \frac{4L\nu_K^2(G)\log 2L}{Kt}$ yields

$$
\mathbf{E}\left[\rho(\boldsymbol{a})\right] \leq \mathbf{E}\left[\rho_c(\boldsymbol{a})\right] \leq \frac{4L\nu_K^2(G)}{Kt}\left(\log 2L + 1 + \frac{Kt}{4} e^{-\frac{K(1-t)^2}{4}}\right). \tag{40}
$$

Let $g(K, t) := \frac{Kt}{4} e^{-\frac{K(1-t)^2}{4}}$. It is trivial to show that $g(K, 0.5) \leq 1$ for all K (for $t = 0.5$, $\max_K g(K, 0.5) = \frac{2}{e}$). Therefore, $\mathbf{E}\left[\rho(\boldsymbol{a})\right] \leq \frac{8L\nu_K^2(G)}{K}(\log 2L + 2)$, which completes the proof. $\blacksquare$

B. Discussion

Remark 1: In Theorem 2, we find an upper bound on $\mathbf{E}\left[\rho(\boldsymbol{a})\right]$ by finding an upper bound on $\mathbf{E}\left[\rho_c(\boldsymbol{a})\right]$ and using the fact that for all vectors $\boldsymbol{a}$, we have $\rho(\boldsymbol{a}) \leq \rho_c(\boldsymbol{a})$. However, we should note that this inequality gets tighter as M (the number of columns of A) increases. For small M the interlacing technique results in a looser bound.

Remark 2: By taking $G = I_N$ and noting that $\nu_K(I_N) = \sqrt{\frac{K}{L}}$, (40) leads to an upper bound on $\mathbf{E}\left[\rho_c(\boldsymbol{a})\right]$ when the signal $\boldsymbol{a}$ is K-sparse in the time domain (specifically, $\mathbf{E}\left[\rho_c(\boldsymbol{a})\right] \leq 8(\log 2L + 2)$). Although this bound scales logarithmically in the ambient dimension L, it does not show a dependency on the sparsity level K of the vector $\boldsymbol{a}$. Over multiple simulations where we have computed the sample
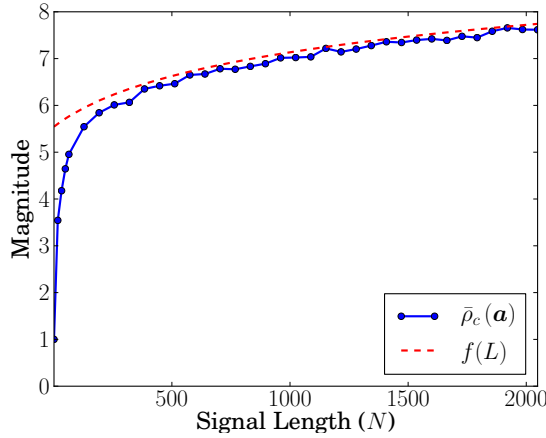
Fig. 4: *Empirical results corresponding to Example 5: Sample mean of $\rho_c(\boldsymbol{a})$ in the time domain for full vectors $\boldsymbol{a} \in \mathbb{R}^N$ where $M = 256$ is fixed. Also plotted is $f(L) = \log(L)$, where $L = N + M - 1$.*

mean $\bar{\rho}_c(\boldsymbol{a})$, we have observed a linear behavior of the quantity $\frac{K}{\bar{\rho}_c(\boldsymbol{a})}$ as K increases, and this leads us to the conjecture below. Although at this point we are not able to prove the conjecture, the proposed bound matches closely with empirical data.

Conjecture 1: Fix N and M. Let $\boldsymbol{a} \in \mathbb{R}^N$ be a random K-sparse vector whose K non-zero entries (on an arbitrary support S) are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \frac{1}{K})$. Then

$$\mathbf{E}\left[\rho_c(\boldsymbol{a})\right] \sim \frac{K}{c_1 K + c_2},$$

where $c_1 = \frac{1}{c \log L}$ for some constant c, and $c_2 = 1 - c_1$.

The conjecture follows from our empirical observation that $\frac{K}{\bar{\rho}_c(\boldsymbol{a})} \sim c_1 K + c_2$ for some constants c_1 and c_2, the fact that $\rho_c(\boldsymbol{a}) = 1$ for $K = 1$, and the observation that $\bar{\rho}_c(\boldsymbol{a}) \sim c \log L$ when $K = N$ for large N. In the following examples, we illustrate these points and show how the conjectured bound can sharply approximate the empirical mean of $\rho_c(\boldsymbol{a})$.

Example 5: In this experiment, we fix $M = 256$ and take $G = I_N$. For each value of N, we construct 1000 random non-sparse vectors $\boldsymbol{a} \in \mathbb{R}^N$ whose N entries are drawn from a Gaussian distribution with mean zero and variance $\frac{1}{N}$. We let $\bar{\rho}_c(\boldsymbol{a})$ denote the sample mean of $\rho_c(\boldsymbol{a})$ across these 1000 signals. The results, as a function of N, are plotted in Fig. 4. Also plotted is the function $f(L) = \log(L)$ where $L = N + M - 1$; this closely approximates the empirical data.

Example 6: In this experiment, we fix $N = 1024$. For each value of K, we construct 1000 random sparse vectors $\boldsymbol{a} \in \mathbb{R}^N$ with random support and having K non-zero entries drawn from a Gaussian
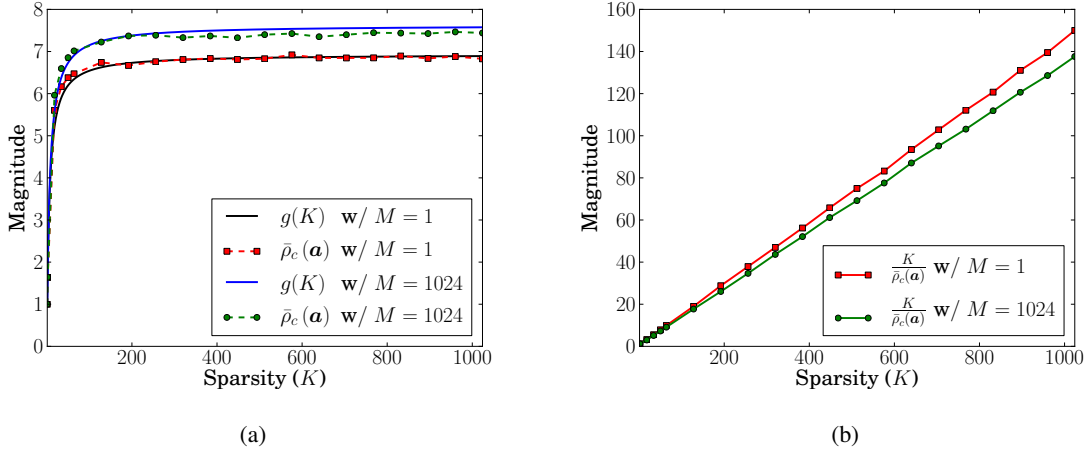
Fig. 5: *Empirical results corresponding to Example 6. (a) Simulation results vs. the conjectured bound $g(K) = \frac{K}{c_1 K + c_2}$ with $c = 1$. (b) Linearity of $\frac{K}{\bar{\rho}_c(\boldsymbol{a})}$.*

distribution with mean zero and variance $\frac{1}{K}$. We let $\bar{\rho}_c(\boldsymbol{a})$ denote the sample mean of $\rho_c(\boldsymbol{a})$ across these 1000 signals. The results, as a function of K for two fixed values $M = 1$ and $M = 1024$, are plotted in Fig. 5.

Remark 3: As a final note in this section, the result of Theorem 2 can be easily extended to the case when $G \in \mathbb{C}^{N \times N}$ is a complex orthobasis and $\boldsymbol{q}$ and $\boldsymbol{a}$ are complex vectors. The bounds can be derived in a similar way and we do not state them for the sake of saving space.

C. A Quadratic RIP Bound and Non-uniform Recovery

An approach identical to the one taken by Baraniuk et al. [7] can be used to establish the RIP for Toeplitz matrices X based on the CoM inequalities given in Theorem 1. As mentioned in Section II, the bounds of the CoM inequalities for Toeplitz matrices are looser by a factor of $2\rho(\boldsymbol{a})$ or $2\mu(\boldsymbol{a})$ as compared to the ones for unstructured X. Since $\rho(\boldsymbol{a})$ is bounded by K for all K-sparse signals in the time domain (the deterministic bound), with straightforward calculations a quadratic estimate of the number of measurements in terms of sparsity ($M \sim K^2$) can be achieved for Toeplitz matrices. As mentioned earlier, on the other hand, there exists an extremely non-uniform distribution of $\rho(\boldsymbol{a})$ over the set of all K-sparse signals $\boldsymbol{a}$, for as Theorem 2 states, if a random model is imposed on $\boldsymbol{a}$, an upper bound on the typical value of $\rho(\boldsymbol{a})$ scales logarithmically in the ambient dimension L. This suggests that for most K-sparse signals $\boldsymbol{a}$ the value of $\rho(\boldsymbol{a})$ is much smaller than K (observe that $8(\log 2L + 2) \ll K$ for many

$$\{x_k\}_{k=1}^{N+M-1} \longrightarrow \boxed{\{a_k\}_{k=1}^{N}} \xrightarrow{\;+\;} \boldsymbol{y} = X\boldsymbol{a} + \boldsymbol{z}$$
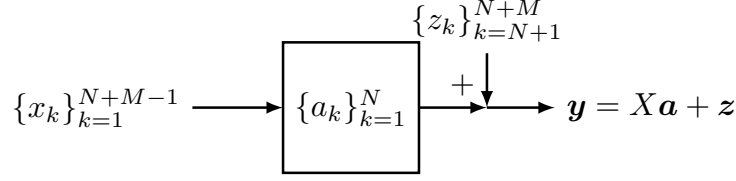
with $\{z_k\}_{k=N+1}^{N+M}$ added.

Fig. 6: *FIR filter of order N with impulse response $\{a_k\}_{k=1}^{N}$.*

signals of practical interest). Only for a very small set of signals does the value of $\rho(\boldsymbol{a})$ approach the deterministic bound of K. One can show, for example, that for any K-sparse signal whose K non-zero entries are all the same, we have $\rho(\boldsymbol{a}) \leq \rho_c(\boldsymbol{a}) = K$ (Example 4). This non-uniformity of $\rho(\boldsymbol{a})$ over the set of sparse signals may be useful for proving a non-uniform recovery bound or for strengthening the RIP result; our work on these fronts remains in progress. Using different techniques than pursued in the present paper (non-commutative Khintchine type inequalities), a non-uniform recovery result with a linear estimate of M in K up to log-factors has been proven by Rauhut [9]. For a detailed description of non-uniform recovery and its comparison to uniform recovery, one could refer to a paper by Rauhut [Sections 3.1 and 4.2, [19]]. The behavior of $\rho(\boldsymbol{a})$ also has important implications in the binary detection problem which we discuss in the next section.

V. COMPRESSIVE BINARY DETECTION

A. Problem Setup

In this section, we address the problem of detecting a change in the dynamics of a linear system. We aim to perform the detection from the smallest number of observations, and for this reason, we call this problem Compressive Binary Detection (CBD).

We consider an FIR filter with a known impulse response $\boldsymbol{a} = \{a_k\}_{k=1}^{N}$. The response of this filter to a test signal $\boldsymbol{x} = \{x_k\}_{k=1}^{N+M-1}$ is described in (3). We suppose the output of this filter is corrupted by random additive measurement noise $\boldsymbol{z}$. Fig. 6 shows the schematic of this measurement process.

From a collection of M measurements $\boldsymbol{y}$ with $M < N$, our specific goal is to detect whether the dynamics of the system have changed to a different impulse response $\boldsymbol{b} = \{b_k\}_{k=1}^{N}$, which we also assume to be known. Since the the nominal impulse response $\boldsymbol{a}$ is known, the expected response $X\boldsymbol{a}$ can be subtracted off from $\boldsymbol{y}$, and thus without loss of generality, our detection problem can be stated as follows [26]: Distinguish between two events which we define as $\mathcal{E}_0 \triangleq \{\boldsymbol{y} = \boldsymbol{z}\}$ and $\mathcal{E}_1 \triangleq \{\boldsymbol{y} = X\boldsymbol{c} + \boldsymbol{z}\}$, where $\boldsymbol{c} = \boldsymbol{b} - \boldsymbol{a}$ and $\boldsymbol{z}$ is a vector of i.i.d. Gaussian noise with variance σ^2.

For any detection algorithm, one can define the false-alarm probability $P_{FA} \triangleq \mathbf{P}\{(\mathcal{E}_1 \text{ chosen when } \mathcal{E}_0)\}$ and the detection probability $P_D \triangleq \mathbf{P}\{(\mathcal{E}_1 \text{ chosen when } \mathcal{E}_1)\}$. A Receiver Operating Curve (ROC) is a plot of P_D as a function of P_{FA}. A Neyman-Pearson (NP) detector maximizes P_D for a given limit on the failure probability, $P_{FA} \leq \alpha$. The NP test for our problem can be written as $\boldsymbol{y}^T X \boldsymbol{c} \underset{\mathcal{E}_0}{\overset{\mathcal{E}_1}{\gtrless}} \gamma$, where the threshold γ is chosen to meet the constraint $P_{FA} \leq \alpha$. Consequently, we consider the detection statistic $d := \boldsymbol{y}^T X \boldsymbol{c}$. By evaluating d and comparing to the threshold γ, we are now able to decide between the two events $\mathcal{E}_0$ and $\mathcal{E}_1$. To fix the failure limit, we set $P_{FA} = \alpha$ which leads to

$$P_D(\alpha) = Q\left(Q^{-1}(\alpha) - \frac{\|X\boldsymbol{c}\|_2}{\sigma}\right), \tag{41}$$

where $Q(q) = \frac{1}{\sqrt{2\pi}} \int_q^\infty e^{-\frac{u^2}{2}} \, du$. As is evident from (41), for a given α, $P_D(\alpha)$ directly depends on $\|X\boldsymbol{c}\|_2$. On the other hand, because X is a compressive random Toeplitz matrix, Theorem 1 suggests that $\|X\boldsymbol{c}\|_2$ is concentrated around its expected value with high probability and with a tail probability bound that decays exponentially in M divided by $\rho(\boldsymbol{c})$. Consequently, one could conclude that for fixed M, the behavior of $\rho(\boldsymbol{c})$ affects the behavior of $P_D(\alpha)$ over α. The following example illustrates this dependency.

Example 7: (Detector Performance) Assume with a failure probability of $P_{FA} = \alpha = 0.05$, a detection probability of $P_D(\alpha) = 0.95$ is desired. Assume $\sigma = 0.3$. From (41) and noting that $Q(-1.6449) = 0.95$ and $Q^{-1}(0.05) = 1.6449$, one concludes that in order to achieve the desired detection, $\|X\boldsymbol{c}\|_2$ should exceed $0.3 \times 2 \times 1.6449 = 0.9869$ (i.e., $\|X\boldsymbol{c}\|_2^2 \geq 0.9741$). On the other hand, for a Toeplitz X with i.i.d. entries drawn from $\mathcal{N}\left(0, \frac{1}{M}\right)$, $\mathbf{E}\left[\|X\boldsymbol{c}\|_2^2\right] = \|\boldsymbol{c}\|_2^2$. Assume without loss of generality, $\|\boldsymbol{c}\|_2 = 1$. Thus, from Theorem 1 and the bound in (8), we have for $\epsilon \in (0, 1)$

$$\mathbf{P}\left\{\|X\boldsymbol{c}\|_2^2 - 1 \leq -\epsilon\right\} \leq e^{-\frac{\epsilon^2 M}{8\mu(\boldsymbol{c})}} \leq e^{-\frac{\epsilon^2 M}{8\rho(\boldsymbol{c})}}. \tag{42}$$

Therefore, for a choice of $\epsilon = 1 - 0.9741 = 0.0259$ and from (42), one could conclude that

$$\mathbf{P}\left\{\|X\boldsymbol{c}\|_2^2 \leq 0.9741\right\} \leq e^{-\frac{6.7 \times 10^{-4} M}{8\rho(\boldsymbol{c})}}.$$

Consequently, for $\zeta \in (0, 1)$, if $M \geq \frac{16\rho(\boldsymbol{c})}{6.7 \times 10^{-4}} \log \zeta^{-1}$, then with probability at least $1 - \zeta^2$, $\|X\boldsymbol{c}\|_2^2$ exceeds 0.9741, achieving the desired detection performance. Apparently, M depends on $\rho(\boldsymbol{c})$ and qualitatively, one could conclude that for a fixed M, a signal $\boldsymbol{c}$ with small $\rho(\boldsymbol{c})$ leads to better detection (i.e., maximized $P_D(\alpha)$ over α). Similarly, a signal $\boldsymbol{c}$ with large $\rho(\boldsymbol{c})$ is more difficult to reliably detect.

In the next section, we examine signals of different $\rho(\boldsymbol{c})$ values and show how their ROCs change. It is interesting to note that this dependence would not occur if the matrix X were unstructured (which, of
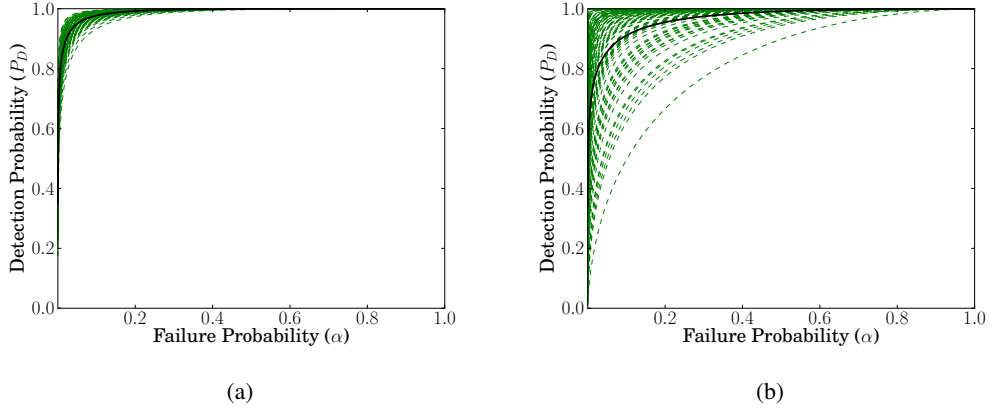
Fig. 7: *ROCs for* 1000 *random matrices* X *for a fixed signal* c *with* $\rho(c) = 45.6$. *(a) Unstructured* X. *(b) Toeplitz* X. *The solid black curve is the average of* 1000 *curves.*

course, would not apply to the convolution-based measurement scenario considered here but is a useful comparison) as the CoM behavior of unstructured Gaussian matrices is agnostic to the signal c.

B. Empirical Results and ROCs

In several simulations, we examine the impact of $\rho(c)$ on the detector performance. To begin, we fix a signal $c \in \mathbb{R}^{256}$ with 50 non-zero entries all taking the same value; this signal has $\|c\|_2 = 1$ and $\rho(c) = 45.6$ with our choice of $M = 128$. We generate 1000 random unstructured and Toeplitz matrices X with i.i.d. entries drawn from $\mathcal{N}\left(0, \frac{1}{M}\right)$. For each matrix X, we compute a curve of P_D over P_{FA} using (41); we set $\sigma = 0.3$. Figures 7(a) and 7(b) show the ROCs resulting from the unstructured and Toeplitz matrices, respectively. As can be seen, the ROCs associated with Toeplitz matrices are more scattered than the ROCs associated with unstructured matrices. This is in fact due to the weaker concentration of $\|Xc\|_2$ around its expected value for Toeplitz X (recall (7) and (8)) as compared to unstructured X (recall (9)).

To compare the ROCs among signals having different $\rho(c)$ values, we design a simulation with 6 different signals. Each signal again has $\|c\|_2 = 1$, and we take $\sigma = 0.3$ as above. Figures 8(a) and 8(b) plot the average ROC for each signal over 1000 random unstructured and Toeplitz matrices, respectively. Two things are evident from these plots. First, the plots associated with Toeplitz matrices show a signal dependency while the ones associated with unstructured matrices are signal-agnostic. Second, with regards to the plots associated with Toeplitz X, we see a decrease in the curves (i.e., inferior detector performance)
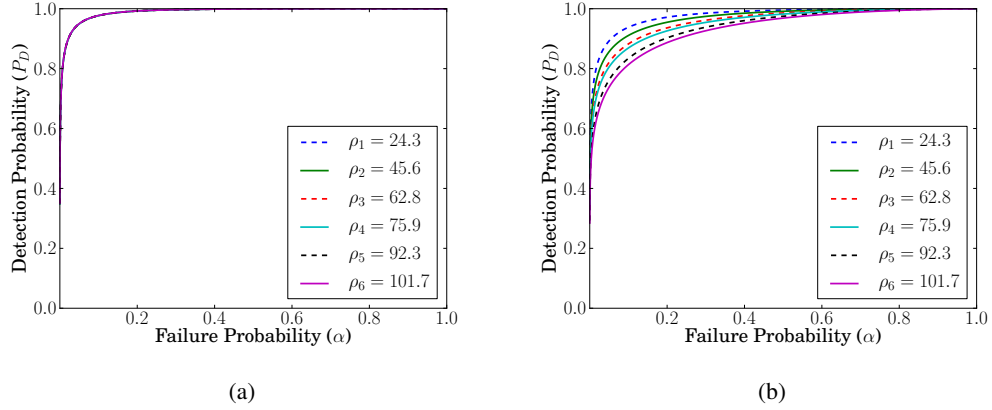
Fig. 8: *Average ROCs over* 1000 *random matrices* X *for 6 different signals* c. *(a) Unstructured* X. *All curves are overlapping. (b) Toeplitz* X. *The curves descend in the same order they appear in legend box.*

for signals with larger values of $\rho\left(c\right)$.

In summary, our theory suggests and our simulations confirm that the value of $\rho\left(c\right)$ has a direct influence on the detector performance. From a systems perspective, as an example, this means that detecting changes in systems having a sparse impulse response in the time domain (e.g., communication channels with multipath propagation) will be easier than doing so for systems having a sparse impulse response in the frequency domain (e.g., certain resonant systems). It is worth mentioning that while detection analysis of systems with sparse impulse response is interesting, our analysis can be applied to situations where neither the impulse responses a and b nor the change c are sparse.

ACKNOWLEDGMENT

APPENDIX

A. Proof of Lemma 5

Proof We start by proving a more general version of (35). Let $z_1, z_2, \ldots, z_n$ be random variables. Consider the event $\mathcal{E}_A \triangleq \{z_1 < c_1 U \text{ and } z_2 < c_2 U \text{ and } \cdots z_n < c_n U\}$ where $c_1, c_2, \ldots, c_n$ are fixed numbers that sum to 1. It is trivial to see that if $\mathcal{E}_A$ happens, then the event $\mathcal{E}_B \triangleq \{z_1 + z_2 + \cdots + z_n < U\}$

must also occur. Consequently, $\mathbf{P}\left\{(\mathcal{E}_B)^c\right\} \leq \mathbf{P}\left\{(\mathcal{E}_A)^c\right\}$, where

$$(\mathcal{E}_A)^c = \{z_1 \geq c_1 U \ \text{or}\ z_2 \geq c_2 U \ \text{or}\ \cdots \ z_n \geq c_n U\}.$$

Using the union bound, we have $\mathbf{P}\left\{(\mathcal{E}_A)^c\right\} \leq \mathbf{P}\left\{z_1 \geq c_1 U\right\} + \mathbf{P}\left\{z_2 \geq c_2 U\right\} + \cdots + \mathbf{P}\left\{z_n \geq c_n U\right\}$ which completes the proof. The inequality (35) is a special case of this result with $c_1 = c_2 = 0.5$.

We follow a similar approach for proving (36) where z_1 and z_2 are positive random variables. Consider the event $\mathcal{E}_A \triangleq \{z_1 < U_1 \ \text{and}\ z_2 > U_2\}$. If $\mathcal{E}_A$ occurs, then $\mathcal{E}_B \triangleq \left\{\frac{z_1}{z_2} < \frac{U_1}{U_2}\right\}$ must also occur. Consequently, $\mathbf{P}\left\{(\mathcal{E}_B)^c\right\} \leq \mathbf{P}\left\{(\mathcal{E}_A)^c\right\}$, where $(\mathcal{E}_A)^c = \{z_1 \geq U_1 \ \text{or}\ z_2 \leq U_2\}$. Using the union bound, we have $\mathbf{P}\left\{(\mathcal{E}_A)^c\right\} \leq \mathbf{P}\left\{z_1 \geq U_1\right\} + \mathbf{P}\left\{z_2 \leq U_2\right\}$. ■

B. Proof of Proposition 2

Before proving Proposition 2, we state the following lemma.

Lemma 6: (*Hoeffding's inequality for real-valued Gaussian sums*) Let $\boldsymbol{b} \in \mathbb{R}^N$ be fixed, and let $\boldsymbol{\epsilon} \in \mathbb{R}^N$ be a random vector whose N entries are i.i.d. random variables drawn from a Gaussian distribution with $\mathcal{N}(0, \sigma^2)$. Then, for any $u > 0$,

$$\mathbf{P}\left\{\left|\sum_{i=1}^N \epsilon_i b_i\right| \geq u\right\} \leq e^{-\frac{u^2}{2\sigma^2 \|\boldsymbol{b}\|_2^2}}.$$

Proof First note that the random variable $\sum_{i=1}^N \epsilon_i b_i$ is also Gaussian with distribution $\mathcal{N}\left(0, \sigma^2 \|\boldsymbol{b}\|_2^2\right)$. Applying a Gaussian tail bound to this distribution yields the inequality [27]. ■

Using the result of Lemma 6, we can complete the proof of Proposition 2.

Proof Let $b_i = r_i + q_i j \ \forall i$ where r_i is the real part of b_i and q_i is the imaginary part. Then we have

$$\mathbf{P}\left\{\left|\sum_{i=1}^N \epsilon_i b_i\right| \geq \|\boldsymbol{b}\|_2 u\right\} = \mathbf{P}\left\{\left|\left(\sum_{i=1}^N \epsilon_i r_i\right) + \left(\sum_{i=1}^N \epsilon_i q_i\right) j\right| \geq \|\boldsymbol{b}\|_2 u\right\}$$

$$= \mathbf{P}\left\{\left|\left(\sum_{i=1}^N \epsilon_i r_i\right) + \left(\sum_{i=1}^N \epsilon_i q_i\right) j\right|^2 \geq \|\boldsymbol{b}\|_2^2 u^2\right\}$$

$$\leq \mathbf{P}\left\{\left(\sum_{i=1}^N \epsilon_i r_i\right)^2 \geq \frac{\|\boldsymbol{b}\|_2^2 u^2}{2}\right\} + \mathbf{P}\left\{\left(\sum_{i=1}^N \epsilon_i q_i\right)^2 \geq \frac{\|\boldsymbol{b}\|_2^2 u^2}{2}\right\}$$

$$= \mathbf{P}\left\{\left|\sum_{i=1}^N \epsilon_i r_i\right| \geq \frac{\|\boldsymbol{b}\|_2 u}{\sqrt{2}}\right\} + \mathbf{P}\left\{\left|\sum_{i=1}^N \epsilon_i q_i\right| \geq \frac{\|\boldsymbol{b}\|_2 u}{\sqrt{2}}\right\}$$

$$\leq e^{-\frac{\|\boldsymbol{b}\|_2^2 u^2}{4\|\boldsymbol{r}\|_2^2 \sigma^2}} + e^{-\frac{\|\boldsymbol{b}\|_2^2 u^2}{4\|\boldsymbol{q}\|_2^2 \sigma^2}} \leq 2e^{-\frac{u^2}{4\sigma^2}},$$

where the first inequality uses Lemma 5 and the last inequality uses Lemma 6 and the facts that $\frac{\|\boldsymbol{b}\|_2}{\|\boldsymbol{r}\|_2} \geq 1$ and $\frac{\|\boldsymbol{b}\|_2}{\|\boldsymbol{q}\|_2} \geq 1$. ∎

REFERENCES

[1] M. Ledoux, *The concentration of measure phenomenon*. Amer Mathematical Society, 2001.

[2] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[3] D. Achlioptas, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *Journal of Computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.

[4] W. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a Hilbert space," *Contemporary mathematics*, vol. 26, pp. 189–206, 1984.

[5] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[6] E. Candès and T. Tao, "Decoding via linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[7] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[8] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Uniform uncertainty principle for Bernoulli and subgaussian ensembles," *Constructive Approximation*, vol. 28, no. 3, pp. 277–289, 2008.

[9] H. Rauhut, "Circulant and Toeplitz matrices in compressed sensing," *Proc. SPARS*, vol. 9, 2009.

[10] J. Haupt, W. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5862–5875, 2010.

[11] J. Romberg, "Compressive sensing by random convolution," *SIAM Journal on Imaging Sciences*, vol. 2, no. 4, pp. 1098–1128, 2009.

[12] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin, "Concentration of measure inequalities for compressive Toeplitz matrices with applications to detection and system identification," *Proc. of 49th IEEE Conf. on Decision and Control*, pp. 2922–2929, 2010.

[13] J. Y. Park, H. L. Yap, C. J. Rozell, and M. B. Wakin, "Concentration of measure for block diagonal matrices with applications to compressive signal processing," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5859–5875, 2011.

[14] M. B. Wakin, B. M. Sanandaji, and T. L. Vincent, "On the observability of linear systems from random, compressive measurements," *Proc. of 49th IEEE Conf. on Decision and Control*, pp. 4447–4454, 2010.

[15] W. Bajwa, J. Haupt, G. Raz, and R. Nowak, "Compressed channel sensing," *Proc. of 42nd Annual Conf. on Inform. Sciences and Systems (CISS08)*, pp. 5–10, 2008.

[16] B. M. Sanandaji, T. L. Vincent, M. B. Wakin, R. Tóth, and K. Poolla, "Compressive system identification of LTI and LTV ARX models," *Proc. of the 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 791–798, 2011.

[17] J. Tropp, M. Wakin, M. Duarte, D. Baron, and R. Baraniuk, "Random filters for compressive sampling and reconstruction," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 872–875, 2006.

[18] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak, "Toeplitz-structured compressed sensing matrices," *IEEE/SP* 14*th Workshop on Statistical Signal Processing*, pp. 294–298, 2007.

[19] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical Foundations and Numerical Methods for Sparse Recovery*, vol. 9, pp. 1–92, 2010.

[20] H. Rauhut, J. Romberg, and J. A. Tropp, "Restricted isometries for partial random circulant matrices," *Appl. Comput. Harmon. Anal.*, 2011.

[21] F. Krahmer, S. Mendelson, and H. Rauhut, "Suprema of chaos processes and the restricted isometry property," 2012.

[22] H. L. Yap and C. J. Rozell, "On the relation between block diagonal matrices and compressive Toeplitz matrices," October 2011, technical Report.

[23] S. Hwang, "Cauchy's interlace theorem for eigenvalues of Hermitian matrices," *The American Mathematical Monthly*, vol. 111, no. 2, pp. 157–159, 2004.

[24] G. Lugosi, "Concentration-of-measure inequalities," *Lecture Notes*, 2004.

[25] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*.   Springer, 1991.

[26] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, "Signal processing with compressive measurements," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 445–460, 2010.

[27] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.